UTS
UNIVERSITY OF TECHNOLOGY SYDNEY

Minimum-risk Sequence Alignment for the Alignment and Recognition of Action Videos

by

Zhen Wang

A thesis submitted in partial fulfilment for the
degree of Doctor of Philosophy

at the
School of Electrical and Data Engineering
Faculty of Engineering and Information Technology
University of Technology Sydney

July 2018

Declaration of Authorship

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This research was supported by an Australian Government Research Training Program (RTP).

Signed: Production Note: Signature removed prior to publication.

Date: 18/07/2018

Abstract

Temporal alignment of videos is an important requirement of tasks such as video comparison, analysis and classification. In the context of action analysis and action recognition, the main guiding element for the temporal alignment are the human actions depicted in the videos. While well-established alignment algorithms such as dynamic time warping are available, they still heavily rely on basic linear cost models and heuristic parameter tuning. Inspired by the success of the hidden Markov support vector machine for pairwise alignment of protein sequences, in this thesis we present a novel framework which combines the flexibility of a pair hidden Markov model (PHMM) with the effective parameter training of the structural support vector machine (SSVM). The framework extends the scoring function of SSVM to capture the similarity of two input frame sequences and introduces suitable feature and loss functions. During learning, we leverage these loss functions for regularised empirical risk minimisation and effective parameter selection.

We have carried out extensive experiments with the proposed technique (nicknamed as EHMM-SSVM) against state-of-the-art algorithms such as dynamic time warping (DTW) and generalized canonical time warping (GCTW) on pairs of human actions from four well-known datasets. The results show that the proposed model has been able to outperform the compared algorithms by a large margin in terms of alignment accuracy.

In the second part of this thesis we employ our alignment approach to tackle the task of human action recognition in video. This task is highly challenging due to the substantial variations in motion performance, recording settings and inter-personal differences. Most current research focuses on the extraction of effective features and the design of suitable classifiers. Conversely, in this thesis we tackle this problem by a dissimilarity-based approach where classification is performed in terms of minimum distance from templates and where the distance is based on the score of our alignment model, the EHMM-SSVM. In turn, the templates are chosen by means of prototype selection techniques from the available samples of each class. Experimental results over two popular human action datasets

have showed that the proposed approach has been capable of achieving an accuracy higher than many existing methods and comparable to a state-of-the-art action classification algorithm.

Acknowledgements

I would like to take this opportunity to acknowledge the enormous support from my supervisor, Prof. Massimo Piccardi, the very useful suggestions from my group and the friendly environment of our lab especially during coffees, lunches and dinners.

Contents

List of Figures

List of Tables